OCCWEAVE: STRUCTURE-GUIDED REFINEMENT AND FUSION FOR COOPERATIVE SEMANTIC OCCUPANCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Cooperative semantic occupancy prediction combines observations from connected vehicles to recover dense scene geometry and semantics. Yet collaboration can leave distorted boundaries, missing thin structures, and fragmented local regions in the final occupancy prediction. Preserving these structures requires both informative local features and effective integration of observations from different agents. We propose OccWeave, a framework that introduces visual structural priors to connect local feature refinement with cooperative evidence weighting. Structure-Prior Feature Refinement (SPFR) associates prior features with each agent's local 3D representation and incorporates them through a gated residual. Structure-aware Cooperative Reliability Fusion (SCRF) uses the refined features, prior-induced residuals, cross-agent discrepancies, and relative pose to estimate local source weights. We realize OccWeave through a tri-plane-to-voxel architecture following CoHFF and a Gaussian-based architecture following VOGS-CP. On Semantic-OPV2V, the voxel and Gaussian implementations improve overall mIoU by 4.35 and 2.39 percentage points over their respective baselines. Both implementations also improve semantic prediction on ego-invisible non-empty voxels, in class-boundary bands, and across the thin-structure class group relative to their respective baselines.

1 INTRODUCTION

Semantic occupancy prediction describes scene geometry and semantics on a dense 3D grid, capturing object extent, drivable surfaces, and roadside structures beyond a sparse set of detection boxes (Xu et al., 2025a). Recent camera-based methods recover this representation from images using voxel, tri-plane, or sparse-query features (Cao & de Charette, 2022; Huang et al., 2023; Li et al., 2023a; Wei et al., 2023; Zhang et al., 2023; Tang et al., 2024), supported by large-scale occupancy benchmarks (Tian et al., 2023; Wang et al., 2023). However, limited viewpoints and occlusion leave a single vehicle with incomplete observations of the surrounding scene.

Cooperative perception addresses this limitation through observations from connected autonomous vehicles (CAVs) (Wang et al., 2020; Xu et al., 2022b; Li et al., 2022a). Existing methods explore pose-aware aggregation, graph- and transformer-based fusion, and communication-efficient feature selection (Li et al., 2021; Hu et al., 2022; Xu et al., 2022a; 2023a). For semantic occupancy, compact plane representations, semantic Gaussian primitives, and sparse feature quantization offer different ways to represent and exchange collaborative evidence (Song et al., 2024; Chen et al., 2026; Li et al., 2026). These advances make complementary observations available for dense scene prediction, including regions that are poorly observed by the ego vehicle.

However, access to complementary observations does not by itself ensure structurally faithful occupancy predictions. Local errors can remain at road–sidewalk boundaries, around thin structures, and along continuous objects or roadside surfaces, even after collaborative feature fusion. Figure 1 illustrates these errors in both voxel-based CoHFF and Gaussian-based VOGS-CP predictions: boundary distortion, broken continuity, thin-structure omission, and thin-structure thickening. The remaining challenge is to translate complementary observations into occupancy predictions that preserve object boundaries, thin structures, and local continuity.

Preserving local structure requires informative local features and appropriate source weighting. Local features may provide weak cues about object extent or class transitions, while different viewpoints

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

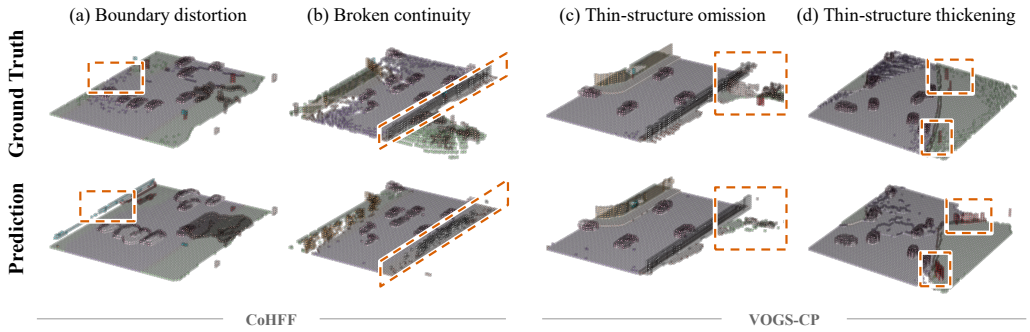

Figure 1: Representative structural errors in cooperative semantic occupancy predictions. The top and bottom rows show ground truth and predictions, respectively. Columns (a–b) show boundary distortion and broken continuity in CoHFF; columns (c–d) show thin-structure omission and thickening in VOGS-CP. Orange dashed boxes mark the corresponding regions of interest.

can supply unequal or conflicting evidence for the same region. Geometric alignment establishes correspondence, but not how much each source should contribute. Visual structural priors supply cues about extent, boundaries, and region organization to refine local 3D features and guide cooperative fusion.

We propose OccWeave, which connects local feature refinement with cooperative source weighting through visual structural priors, instantiated here with dense Segment Anything Model (SAM) embeddings (Kirillov et al., 2023). Structure-Prior Feature Refinement (SPFR) associates spatial priors with agent-wise 3D representations and refines local features through a learned gated residual. Structure-aware Cooperative Reliability Fusion (SCRF) uses the refined features, prior-induced residuals, cross-agent discrepancies, and relative pose to determine local CAV contribution weights. We instantiate this design in CoHFF's tri-plane-to-voxel architecture (Song et al., 2024) and VOGS-CP's Gaussian architecture (Chen et al., 2026). On Semantic-OPV2V, the voxel and Gaussian implementations improve overall mIoU by 4.35 and 2.39 percentage points over their baselines, with gains in class-boundary bands, on ego-invisible non-empty voxels, and across the thin-structure class group.

The main contributions of this paper are:

- We propose OccWeave, a cooperative semantic occupancy framework that introduces visual structural priors to connect local 3D feature refinement with cooperative source weighting, improving predictions in structure-sensitive regions.
- We develop a gated residual feature-refinement mechanism and use the refined features and prior-induced residuals as source-scoring evidence. Together with cross-agent discrepancies and relative pose, this evidence determines local CAV contribution weights.
- Experiments on Semantic-OPV2V show improved overall semantic occupancy accuracy in both voxel and Gaussian implementations, with gains in class-boundary bands, on ego-invisible non-empty voxels, and across the thin-structure class group.

2 Related Work

Semantic occupancy prediction. Semantic scene completion recovers dense geometry and semantics from partial observations (Song et al., 2017; Behley et al., 2019; Roldão et al., 2020; Xia et al., 2023). Related camera-based BEV methods predict future instances (FIERY (Hu et al., 2021)) or support 3D detection and map segmentation (BEVFormer (Li et al., 2022b)). For dense 3D semantic occupancy, camera-based approaches lift image features into 3D (Cao & de Charette, 2022; Wei et al., 2023), using tri-plane, transformer, or sparse-query representations to balance spatial detail and computation (Huang et al., 2023; Zhang et al., 2023; Li et al., 2023a; Tang et al., 2024). Further work explores view transformation, panoptic prediction, and rendering supervision (Li et al., 2023b; Wang et al., 2024; Pan et al., 2024), supported by large-scale occupancy benchmarks (Tian et al.,

2023; Wang et al., 2023). Single-vehicle observations leave occluded and weakly observed regions dependent on limited local evidence.

Cooperative perception. V2X datasets support feature sharing across vehicles and infrastructure (Xu et al., 2022b; Li et al., 2022a; Yu et al., 2022; Xu et al., 2023b; Xiang et al., 2024). Methods address bandwidth (Chen et al., 2019; Liu et al., 2020b;a; Wang et al., 2020), latency, pose noise, and agent heterogeneity (Lei et al., 2022; Lu et al., 2023; 2024). DiscoNet learns collaboration graphs (Li et al., 2021); V2X-ViT and CoBEVT use transformer fusion (Xu et al., 2022a; 2023a); Where2comm selects communicated features through spatial confidence (Hu et al., 2022). These studies establish spatially selective communication and fusion. We use refined local 3D features, prior-induced residuals, and cross-agent discrepancies to determine source contributions for semantic occupancy.

Cooperative semantic occupancy. Recent benchmarks extend cooperation to dense occupancy prediction and forecasting (Wu et al., 2025; Wang et al., 2025), while TSMV uses two-stream multi-vehicle collaboration for semantic scene completion (Li et al., 2024). CoHFF introduces Semantic-OPV2V and combines cooperative tri-plane features with semantic–occupancy task fusion (Song et al., 2024). VOGS-CP fuses sparse semantic Gaussian primitives across vehicles (Chen et al., 2026), and VQSOP uses sparse-aware vector quantization for communication-efficient occupancy prediction (Li et al., 2026). OccWeave connects structural-prior feature refinement with cooperative source weighting through tri-plane-to-voxel and Gaussian instantiations.

Foundation-model priors for 3D perception. SAM supplies class-agnostic visual cues through dense image embeddings (Kirillov et al., 2023). Foundation-model features support open-vocabulary 3D understanding, mapping, and instance segmentation (Peng et al., 2023; Jatavallabhula et al., 2023; Kerr et al., 2023; Takmaz et al., 2023); SA3D, SAM3D, and SAMPro3D associate SAM masks with 3D scenes through projection and cross-view processing (Cen et al., 2025; Yang et al., 2023; Xu et al., 2025b). For single-agent occupancy, SAMOccNet uses SAM-derived visual details to improve semantic occupancy prediction (Tan et al., 2025). Our method uses structural priors to refine each agent's local features and carries the resulting features and prior-induced residuals into cooperative source scoring. Dense SAM embeddings provide the priors in our implementation.

3 METHOD

3.1 PROBLEM FORMULATION

We consider image-based cooperative semantic occupancy prediction with an ego vehicle e and up to N connected vehicles. The agent set is $\mathcal{A} = \{e, 1, \dots, N\}$. Each agent $a \in \mathcal{A}$ observes synchronized multi-view images $\{I_{a,k}\}_{k=1}^{K_a}$ and provides its pose relative to the ego frame. The model predicts ego-centric semantic occupancy logits $\hat{Y} \in \mathbb{R}^{C \times X \times Y \times Z}$, where C is the number of output classes and (X, Y, Z) is the occupancy-grid resolution.

To cover both dense and sparse cooperative representations, we describe an agent's 3D representation as a set of local elements Ω_a. An element $q \in \Omega_a$ has an ego-frame position $x_{a,q} \in \mathbb{R}^3$ and a feature $h_{a,q} \in \mathbb{R}^D$. Its position in agent a's local frame is denoted by $x_{a,q}^{\text{local}}$, with $x_{a,q} = R_{a\to e} x_{a,q}^{\text{local}} + t_{a\to e}$, where $(R_{a\to e}, t_{a\to e})$ is the rigid transform to the ego frame. We use the local-frame coordinate $x_{a,q}^{\text{local}}$ for within-agent visual-prior association, whereas the ego-frame coordinate $x_{a,q}$ is used for cross-agent alignment and matching. For an ego element $q \in \Omega_e$, $\mathcal{N}_i(q) \subseteq \Omega_i$ denotes the aligned elements of CAV i that can contribute to q, and $\mathcal{I}(q) = \{i \mid \mathcal{N}_i(q) \neq \varnothing\}$ is the set of contributing CAV indices. This notation has two concrete realizations: in the voxel instantiation, q is a voxel and $\mathcal{N}_i(q)$ contains the corresponding ego-frame CAV voxel; in the Gaussian instantiation, q is an ego Gaussian anchor and $\mathcal{N}_i(q)$ contains radius-matched Gaussians from CAV i.

3.2 OVERALL ARCHITECTURE

OccWeave connects local feature refinement with cooperative source weighting (Figure 2). SPFR associates visual priors with local 3D elements and refines their features; SCRF uses this evidence

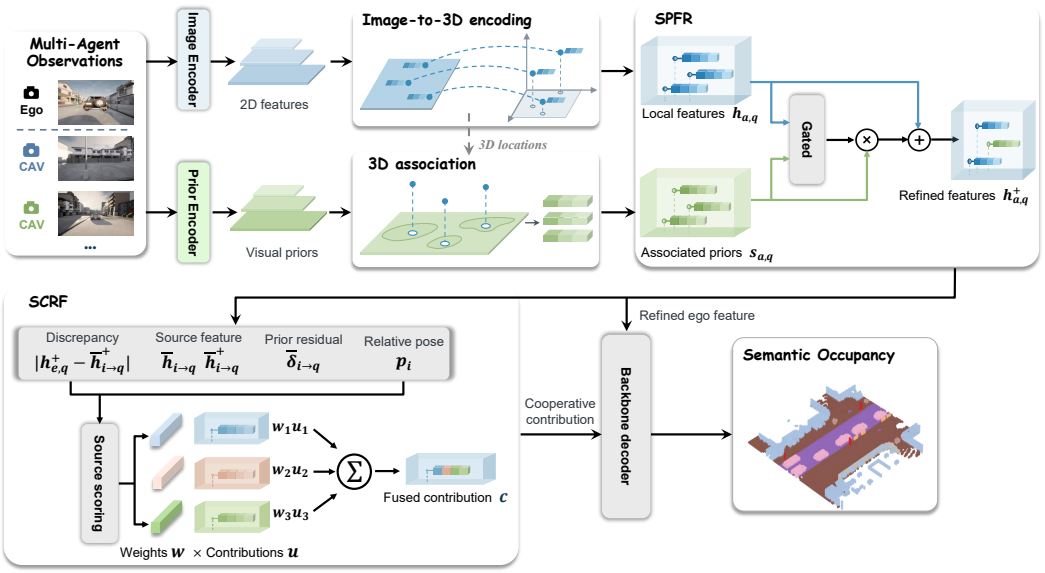

Figure 2: Overview of OccWeave. Backbone-specific encoding produces local 3D features. SPFR refines them with associated structural-prior features, and SCRF weights cooperative contributions using local reliability evidence. The refined ego features and fused contributions are passed to the backbone decoder for semantic occupancy prediction.

to weight CAV contributions. The two instantiations share this computation with separately learned parameters and representation-specific association and contribution operators.

Let P denote the visual prior source and $E^P_{a,k} = \mathcal{E}_P(I_{a,k})$ its spatial feature map for camera k on agent a. The encoder $\mathcal{E}_P$ supplies visual cues that are associated with local 3D elements. Our implementation uses a frozen SAM image encoder (Kirillov et al., 2023). A representation-specific association operator $\mathcal{T}^P_a$ maps the multi-view embeddings to a structural prior at element q:

$$s_{a,q} = \mathcal{T}^P_a\left(\{E^P_{a,k}\}_{k=1}^{K_a}, x^{\text{local}}_{a,q}\right), \qquad s_{a,q} \in \mathbb{R}^D. \tag{1}$$

Let $\mathsf{H}_a = (h_{a,q})_{q\in\Omega_a}$ and $\mathsf{S}_a = (s_{a,q})_{q\in\Omega_a}$ denote the indexed collections of task and associated prior features. The lightweight refinement head ϕ_{ref} uses 3D convolutions to aggregate local spatial context in the voxel implementation and an element-wise MLP in the Gaussian implementation. The refinement gate, prior-induced residual, and refined feature are

$$g_{a,q} = [\sigma(\phi_{\text{ref}}([\mathsf{H}_a, \mathsf{S}_a]))]_q, \qquad \delta_{a,q} = g_{a,q}\, s_{a,q}, \qquad h^+_{a,q} = h_{a,q} + \delta_{a,q}. \tag{2}$$

where $[\cdot,\cdot]$ denotes channel concatenation, $[\cdot]_q$ selects the output at element q, and $g_{a,q} \in [0,1]$ is a scalar gate shared across feature channels. The residual form preserves the task-trained representation while allowing the associated prior evidence to modulate local features.

For each $i \in \mathcal{I}(q)$, a representation-specific pooling operator ρ summarizes the aligned source elements:

$$\begin{aligned} \bar{h}_{i\to q} &= \rho(\{h_{i,q'} : q' \in \mathcal{N}_i(q)\}), \\ \bar{\delta}_{i\to q} &= \rho(\{\delta_{i,q'} : q' \in \mathcal{N}_i(q)\}), \end{aligned} \tag{3}$$

and $\bar{h}^+_{i\to q} = \bar{h}_{i\to q} + \bar{\delta}_{i\to q}$. The reliability evidence is defined consistently as

$$z_{i,q} = \left[h^+_{e,q}, \bar{h}_{i\to q}, \bar{h}^+_{i\to q}, |h^+_{e,q} - \bar{h}^+_{i\to q}|, \bar{\delta}_{i\to q}, p_i\right], \tag{4}$$

where p_i encodes the planar relative pose of CAV i with respect to the ego vehicle: longitudinal displacement, lateral displacement, and yaw difference. Its representation-specific encoding is specified below. Let $\mathcal{Q}_i = \{q \in \Omega_e : i \in \mathcal{I}(q)\}$ and $\mathsf{Z}_i = (z_{i,q})_{q\in\mathcal{Q}_i}$ collect source i's evidence over valid target elements. A score head shared across CAVs produces $r_{i,q} = [\psi_{\text{rel}}(\mathsf{Z}_i)]_q \in \mathbb{R}$, using spatial convolutions for voxels and a per-anchor MLP for Gaussians. Let $u_{i\to q}$ denote the

representation-specific cooperative contribution from CAV i to target element q: a refined voxel feature in the voxel instantiation or a source-grouped Gaussian attribute update in the Gaussian instantiation. With temperature τ, source-normalized weights and the cooperative output are

$$w_{i,q} = \frac{\exp(r_{i,q}/\tau)}{\sum_{j\in\mathcal{I}(q)} \exp(r_{j,q}/\tau)}, \qquad c_q = \sum_{i\in\mathcal{I}(q)} w_{i,q} u_{i\to q}. \tag{5}$$

The refined ego feature provides the scoring reference, while the source-normalized weights determine how strongly the contribution from each available CAV enters the cooperative update.

3.3 Voxel Instantiation

Semantic and structural voxel features. The voxel instantiation reconstructs semantic voxel features through the plane-to-voxel (P2V) operator of the CoHFF architecture. Let T_e^{xy}, T_e^{xz}, and T_e^{yz} denote the ego-associated plane features produced by the CoHFF backbone before the OccWeave voxel-level reliability fusion described below, and let T_i^{xz} and T_i^{yz} denote the corresponding source-specific plane features associated with CAV i. P2V reconstructs the semantic voxel features as

$$\begin{aligned} F_e &= \mathcal{P}_{xy}(T_e^{xy}) + \mathcal{P}_{xz}(T_e^{xz}) + \mathcal{P}_{yz}(T_e^{yz}), \\ F_i &= \mathcal{P}_{xz}(T_i^{xz}) + \mathcal{P}_{yz}(T_i^{yz}), \qquad i = 1, \dots, N, \end{aligned} \tag{6}$$

where $\mathcal{P}_\cdot$ reshapes, permutes, and expands a plane along its missing spatial axis. All $F_a \in \mathbb{R}^{D\times X\times Y\times Z}$ are aligned to the ego grid.

A source-specific projector builds a prior feature pyramid $\{M_{a,k}^\ell\}_{\ell=1}^L = \phi_{\text{prior}}^P(E_{a,k}^P)$. A structural-prior tri-plane encoder uses the semantic branch's geometric lifting interface, and P2V produces ego-aligned structural features S_a. At voxel v, the shared variables are $q = v$, $h_{a,v} = F_a(v)$, and $s_{a,v} = S_a(v)$.

Structure-prior voxel refinement. Equation (2) produces $F_a^+(v) = h_{a,v}^+$ and the prior-induced residual $\delta_{a,v}$. A small 3D convolutional network predicts the scalar gate $A_a(v) = g_{a,v}$, broadcast across feature channels.

Voxel-wise cooperative reliability fusion. Each CAV contributes one feature per aligned voxel, so $\mathcal{N}_i(v)$ is a singleton and ρ selects that feature and its prior-induced residual. For voxel-wise fusion, we use $p_i = p_i^{\text{vox}} = [\Delta x_i, \Delta y_i, \Delta\psi_i]$ and broadcast this pose vector over the spatial grid. A shared 3D convolutional head scores the evidence in Eq. (4). Setting $u_{i\to v} = F_i^+(v)$ in Eq. (5) gives the fused CAV feature $F_{\text{cav}}(v) = c_v$.

Cooperative semantic occupancy decoder. The CoHFF occupancy branch voxelizes ego monocular-depth estimates and maps them to a geometry feature F_{occ} through a compact 3D completion network. Its task-fusion decoder predicts $\hat{Y} = \mathcal{D}_{\text{vox}}([F_{\text{occ}}, F_e^+, F_{\text{cav}}])$.

3.4 Gaussian Instantiation

The Gaussian instantiation follows the VOGS-CP architecture and represents each agent with Gaussian primitives. Let $m \in \Omega_a$ index a Gaussian with center $x_{a,m}$, hidden feature $h_{a,m}$, and physical attributes $\theta_{a,m}$ comprising position, scale, rotation, opacity, and semantics.

Prior-to-Gaussian association and refinement. Before the final single-agent Gaussian refinement, the local-frame center $x_{a,m}^{\text{local}}$ is projected into every camera in which it is visible. Let $\mathcal{V}_{a,m}$ be the set of valid cameras and $\pi_{a,k}$ the projection from the agent frame to camera k. For $|\mathcal{V}_{a,m}| > 0$, bilinear sampling, view averaging, and projection give

$$s_{a,m} = W_{\text{prior}}^P \frac{1}{|\mathcal{V}_{a,m}|} \sum_{k\in\mathcal{V}_{a,m}} \text{Bilinear}\big(E_{a,k}^P, \pi_{a,k}(x_{a,m}^{\text{local}})\big)\,. \tag{7}$$

A scalar MLP gate applies Eq. (2); elements without a valid view receive zero prior features and $\delta_{a,m} = 0$. The refined hidden feature $h_{a,m}^+$ is then consumed by the original single-agent Gaussian

refinement stage of VOGS-CP. In the remainder of this section, $\theta_{a,m}$ denotes the physical Gaussian attributes after this single-agent refinement. The raw feature, refined feature, and prior-induced residual are retained as source-scoring evidence.

Source-grouped Gaussian reliability. CAV Gaussians are rigidly aligned to the ego frame, and radius matching constructs $\mathcal{N}_i(m)$ for each ego anchor m. We instantiate ρ in Eq. (3) as the mean over all matched Gaussians from the same CAV. The pooled features $\bar{h}_{i\to m}$ and $\bar{h}^{+}_{i\to m}$, together with the prior-induced residual $\bar{\delta}_{i\to m}$, form the source-scoring evidence $z_{i,m}$ according to Eq. (4). The relative pose is normalized as $p_i = p_i^{\text{gau}} = [\Delta x_i/40, \Delta y_i/40, \Delta\psi_i/\pi]$, and the Gaussian score head applies Eq. (5) with $\tau = 1$ over $i \in \mathcal{I}(m)$.

For a matched pair (m, n), the pair network predicts $d_{i,m,n} = \phi_{\text{pair}}(\theta_{e,m}, \theta_{i,n})$. We average these updates within each source and apply the corresponding source reliability weights computed for anchor m:

$$u_{i\to m} = \frac{1}{|\mathcal{N}_i(m)|} \sum_{n\in\mathcal{N}_i(m)} d_{i,m,n}, \qquad \Delta\theta_{e,m} = c_m = \sum_{i\in\mathcal{I}(m)} w_{i,m} u_{i\to m}. \tag{8}$$

With mean aggregation, the update refines the ego Gaussian's position, scale, rotation, semantics, and opacity. An anchor with no matched CAV retains its attributes; a single contributing CAV receives a source weight of one. The occupancy head $\mathcal{D}_{\text{gau}}$ from VOGS-CP decodes the refined Gaussian set into $\hat{Y}$.

3.5 Training Objectives

Both instantiations are supervised by ego-centric voxel labels after their representation-specific decoders. Let $y_v^* \in \{1, \dots, C\}$ denote the ground-truth semantic class of voxel v. For class weights α_c, the weighted semantic cross-entropy is

$$\mathcal{L}_{\text{ce}} = -\sum_{v} \sum_{c=1}^{C} \alpha_c \mathbf{1}[y_v^* = c] \log \hat{y}_{v,c}, \tag{9}$$

where $\hat{y}_{v,c} = \text{softmax}_c(\hat{Y}_{:,v})$ is the corresponding predicted probability obtained from the occupancy logits; ignored classes are assigned zero loss weight. We combine it with a Lovász–Softmax term:

$$\mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{lov}}\mathcal{L}_{\text{lov}}. \tag{10}$$

Loss coefficients and the treatment of empty and unlabeled voxels are identical for each baseline and its corresponding OccWeave implementation.

4 Experiments

4.1 Experimental Setup

Dataset and metrics. We train and evaluate OccWeave on Semantic-OPV2V (Xu et al., 2022b; Song et al., 2024), using 1,980 validation frames. Both implementations predict a $100 \times 100 \times 8$ semantic voxel grid over a $40\,\text{m} \times 40\,\text{m}$ horizontal area. Evaluation covers 12 non-empty semantic classes: IoU is the ground-truth class-frequency-weighted average of their class IoUs, and mIoU is their arithmetic mean.

Evaluation protocol. Both the voxel and Gaussian implementations are evaluated on final semantic voxel predictions. Agent counts include the ego vehicle. For BEV evaluation, we project predictions along the height axis and report IoU for Vehicle, Road, and Others; Others comprises the nine classes excluding Vehicle, Road, and Sidewalk. Structural evaluation covers ego-invisible non-empty voxels, class-boundary bands, and a thin-structure class group, defined in Section C. Published reference scores are identified in the result tables.

Implementation details. Structural priors are extracted offline using the frozen SAM ViT-B image encoder (Kirillov et al., 2023) and associated with local 3D features through trainable adapters. The voxel implementation initializes its semantic and completion branches from pretrained models. The Gaussian implementation initializes from a pretrained cooperative model, freezes the single-agent pathway, and fine-tunes the cooperative refiner and the proposed modules. Gaussian checkpoints are selected by validation mIoU and used for subsequent BEV and structural-region evaluation. Both implementations use the objectives in Section 3.5.

4.2 Comparison with Existing Methods

Table 1 compares each implementation with its corresponding baseline in BEV semantic segmentation under two-agent and up-to-seven-agent settings. The voxel instantiation improves vehicle, road, and other-class IoU over CoHFF in both settings. With up to seven agents, these gains are 7.40, 19.27, and 19.15 percentage points, respectively. The Gaussian instantiation achieves road IoU of 87.28% and 90.27%, improving over VOGS-CP by 4.59 and 5.31 points, respectively. Its gains vary across categories: with up to seven agents, vehicle IoU increases by 1.05 points over VOGS-CP, while other-class IoU decreases by 0.94 points. The full BEV comparison, including CoBEVT and VQSOP, is provided in Table A2; the 3D semantic occupancy comparison is reported in Table A1.

Table 1: BEV semantic segmentation IoU (%) for the voxel and Gaussian implementations, with CoHFF (Song et al., 2024) and VOGS-CP (Chen et al., 2026) as their respective baselines. Δ denotes the change in percentage points; bold marks the higher IoU within each baseline–OccWeave pair under the same agent setting.

Implementation	Method	2 agents			Up to 7 agents		
		Vehicle	Road	Others	Vehicle	Road	Others
Voxel	Baseline	47.40	63.36	40.27	64.44	57.28	45.89
	OccWeave	**52.73**	**74.71**	**66.95**	**71.84**	**76.55**	**65.04**
	Δ	+5.33	+11.35	+26.68	+7.40	+19.27	+19.15
Gaussian	Baseline	**70.25**	82.69	**79.37**	75.30	84.96	**80.19**
	OccWeave	70.12	**87.28**	78.55	**76.35**	**90.27**	79.25
	Δ	-0.13	+4.59	-0.82	+1.05	+5.31	-0.94

4.3 Module and Reliability Ablations

Table 2 evaluates local refinement and cooperative fusion in the voxel and Gaussian implementations. SPFR raises mIoU from 34.16% to 36.67% in the voxel implementation and from 37.44% to 38.64% in the Gaussian implementation. With SCRF, the two implementations further reach 38.51% and 39.83%. Both representations follow the same progression: local feature refinement improves prediction, and changing how cooperative features are fused yields further gains. We next distinguish the effects of the aggregation architecture and source scoring.

The SPFR configuration in Table 3 uses the original concatenation-based fusion; Uniform and the learned scoring variants aggregate the same refined CAV features. Switching from concatenation to uniform aggregation improves overall mIoU, but boundary mIoU changes from 30.25% to 30.16%. Full SCRF, using the same aggregation structure as Uniform, raises boundary mIoU to 32.06% and exceeds Uniform by 0.69 points overall. The boundary gain appears with learned source weights; changing the aggregation structure alone yields little change on this subset.

Pre-refinement scoring replaces the evidence in Eq. (4) with ego and source features before SPFR, their cross-agent discrepancy, and relative pose. The resulting weights aggregate the same SPFR-refined features. Compared with Uniform, it improves boundary and ego-invisible mIoU by 0.41 and 0.23 percentage points, respectively, while reducing overall mIoU by 0.38 points. Learning source weights from these inputs thus yields regional gains without an overall improvement. The variant without the explicit prior-induced residual instead gives the scoring head access to refined features and their discrepancy. Both variants fuse the same features and omit the explicit residual input. With refinement-related scoring inputs, overall mIoU increases from 37.44% to 38.13%, and boundary

Table 2: Module ablation on Semantic-OPV2V. Full includes SPFR and SCRF. Scores are percentages; bold marks the best result within each representation. Percentages beside class names indicate class frequencies.

Evaluation item	Voxel representation			Gaussian representation		
Ablation setting	CoHFF	+SPFR	Full	VOGS-CP	+SPFR	Full
IoU	50.46	55.36	**56.67**	72.44	73.01	**73.13**
mIoU	34.16	36.67	**38.51**	37.44	38.64	**39.83**
Building (5.40%)	25.72	26.78	**27.58**	9.61	**13.52**	13.43
Fence (0.85%)	27.83	36.45	**42.06**	29.20	31.49	**34.46**
Terrain (4.80%)	**48.30**	40.12	41.29	**74.51**	74.28	74.39
Pole (0.39%)	42.74	45.89	**48.72**	12.19	14.24	**15.68**
Road (40.53%)	61.77	72.34	**74.33**	83.05	**83.42**	83.30
Sidewalk (35.64%)	**39.62**	39.51	39.42	78.22	78.60	**78.67**
Vegetation (1.11%)	20.59	20.96	**23.15**	20.43	21.15	**21.42**
Vehicles (9.14%)	63.28	71.82	**74.65**	60.49	60.72	**60.83**
Wall (2.01%)	58.27	64.15	**68.94**	36.45	38.09	**42.94**
Guard rail (0.04%)	**1.94**	1.52	1.60	32.50	35.60	**36.41**
Traffic sign (0.05%)	16.33	17.63	**18.37**	8.26	10.68	**14.93**
Bridge (0.04%)	**3.53**	2.91	1.98	**4.35**	1.89	1.45

mIoU from 30.57% to 31.88%. This comparison shows that prior-refined representations also help determine source weights when the fused features are held fixed.

Explicitly supplying the prior-induced residual further raises overall mIoU from 38.13% to 38.51% and boundary mIoU from 31.88% to 32.06%. The scoring head already receives the original and refined features; exposing their difference explicitly helps it use the prior-induced feature change for source scoring. Removing the explicit cross-agent discrepancy reduces boundary mIoU from 32.06% to 30.62% and ego-invisible mIoU from 35.35% to 34.57%, indicating that an explicit ego–CAV feature comparison helps prediction in both subsets. By comparison, removing pose changes overall and boundary mIoU by just 0.04 and 0.06 points: the full and no-pose variants perform similarly in this setting. Together, these comparisons support the connection between SPFR and SCRF: prior-refined features serve as both fusion values and source-scoring evidence, with explicit prior-induced residual and cross-agent discrepancy inputs further improving prediction. The prior-source comparison in Section B further shows that SAM achieves the best overall and regional scores among the evaluated sources in the voxel implementation, supporting its use as the visual prior.

Table 3: SCRF ablation in the voxel implementation with SPFR. Overall: semantic occupancy IoU and mIoU; hard regions: mIoU. Res.: prior-induced residual; Diff.: cross-agent feature discrepancy; in this column, Pre: pre-refinement, ✓: refined. Pre-refinement scoring weights refined fusion values using pre-refinement features and their discrepancies.

Method	Reliability evidence				Overall		Hard regions mIoU	
	Learned	Pose	Res.	Diff.	IoU	mIoU	Ego-invis.	Boundary
+SPFR	–	–	–	–	55.36	36.67	33.75	30.25
+SPFR + Uniform CAV fusion	–	–	–	–	55.92	37.82	34.45	30.16
+SPFR + SCRF with pre-refinement scoring	✓	✓	–	Pre	55.86	37.44	34.68	30.57
+SPFR + SCRF w/o pose	✓	–	✓	✓	56.62	38.47	35.32	32.00
+SPFR + SCRF w/o prior residual	✓	✓	–	✓	56.25	38.13	35.26	31.88
+SPFR + SCRF w/o ego-CAV diff.	✓	✓	✓	–	56.36	38.02	34.57	30.62
+SPFR + SCRF	✓	✓	✓	✓	56.67	38.51	35.35	32.06

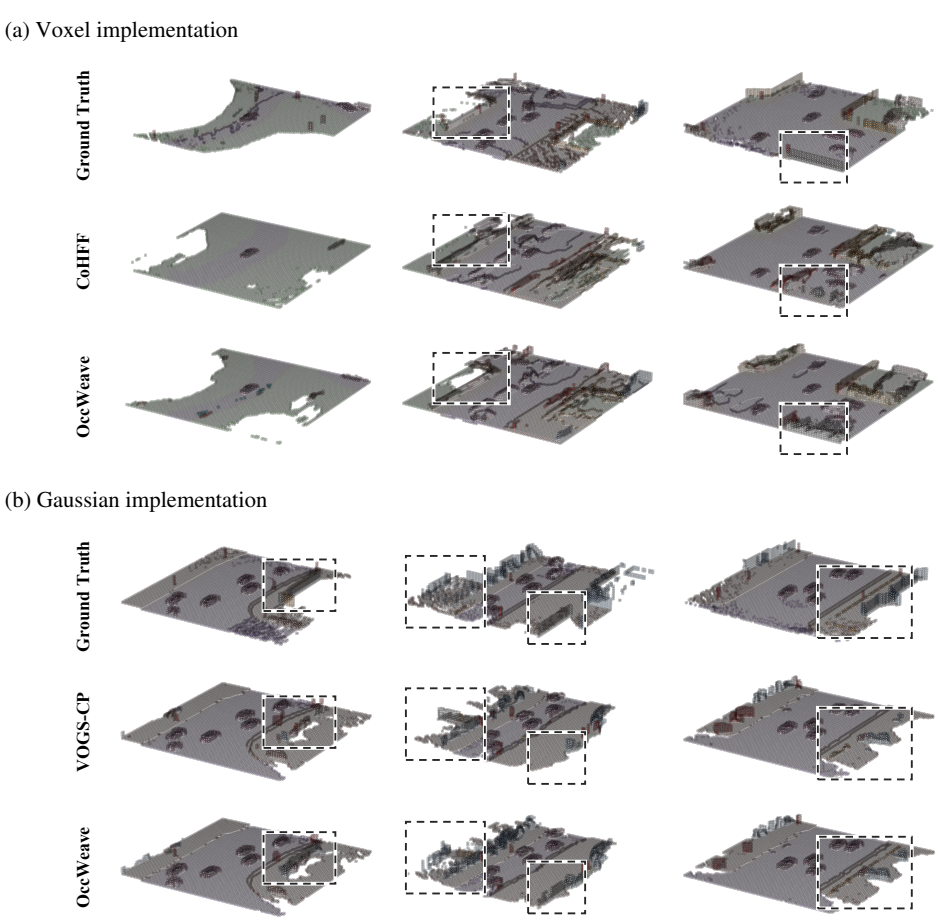

Figure 3: Qualitative semantic occupancy examples on Semantic-OPV2V. The upper and lower groups show the voxel and Gaussian implementations, respectively. Rows show ground truth, baseline, and OccWeave; dashed boxes indicate selected regions.

4.4 Visual Analysis

Figure 3 compares both implementations with their baselines and ground truth. The voxel examples show a curved road closer to the ground-truth outline (left), clearer separation of a narrow roadside structure from its adjacent surface (middle), and less protruding clutter around an elongated structure (right). The Gaussian examples show fewer interruptions in the highlighted strip (right) and more coherent vertical structures and adjoining surfaces (middle). These examples show improved local contours, structure separation, and continuity. Quantitative evaluation also shows gains in class-boundary bands and the thin-structure class group (Table A4). Both implementations also improve on ego-invisible non-empty voxels, which are occupied in cooperative labels but empty in ego-only labels; Section C provides subset definitions and quantitative results.

5 Conclusion

We presented OccWeave, a cooperative semantic occupancy framework that uses visual structural priors to connect local 3D feature refinement with source-specific cooperative weighting. Its voxel and Gaussian instantiations use prior-refined features, prior-induced residuals, and cross-agent discrepancies to guide source scoring in their respective representations. Both implementations improve final semantic occupancy accuracy overall and on the evaluated structure-sensitive subsets and class group, although some rare classes remain difficult. Future work will examine structural-cue reliability under degraded visibility and the cost of extracting dense visual priors.

AI USE STATEMENT

We used GPT-5.6 to assist with drafting sections of the manuscript, editing for readability, translating text, creating and editing code for LaTeX typesetting and table formatting, searching for relevant literature, and formatting references.

The authors reviewed and revised the AI-assisted text and translations. We manually checked the bibliographic metadata of all 51 cited publications, including complete author lists and author order, publication years, volume and issue information, and page ranges or article numbers, where applicable. AI-assisted code was limited to LaTeX typesetting and table formatting. The authors checked its output by inspecting the compiled PDF for layout and table-formatting issues. The authors take responsibility for the final content of this work, including its claims, references, code, and reported results.

REFERENCES

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9297–9307, 2019.

Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3991–4001, 2022.

Jiazhong Cen, Jiemin Fang, Zanwei Zhou, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment anything in 3d with radiance fields. *International Journal of Computer Vision*, 133(8):5138–5160, 2025.

Cheng Chen, Hao Huang, and Saurabh Bagchi. Vision-only gaussian splatting for collaborative semantic occupancy prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (4):2796–2804, 2026. doi: 10.1609/aaai.v40i4.37269.

Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *ACM/IEEE Symposium on Edge Computing*, pp. 88–100, 2019. doi: 10.1145/3318216.3363300.

Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. FIERY: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15273–15282, 2021.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4874–4886, 2022.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9223–9232, 2023.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Varma Keetha, Ayush Tewari, Joshua Tenenbaum, Celso de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set multimodal 3d mapping. In *Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.066.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19672–19682, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick.

Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Zixing Lei, Shunli Ren, Yue Hu, Wenjun Zhang, and Siheng Chen. Latency-aware collaborative perception. In *European Conference on Computer Vision*, volume 13692 of *Lecture Notes in Computer Science*, pp. 316–332, 2022.

Feng Li, Chaokun Zhang, and Gong Chen. Sparse-aware vector quantization for bandwidth-efficient collaborative 3d semantic occupancy prediction. *arXiv preprint arXiv:2607.01928*, 2026.

Junxuan Li, Yuanfang Zhang, Jiayi Han, Peng Han, and Kaiqing Luo. Semantic scene completion in autonomous driving: A two-stream multi-vehicle collaboration approach. *Sensors*, 24(23):7702, 2024.

Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29541–29552, 2021.

Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022a.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9087–9098, 2023a.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, volume 13669 of *Lecture Notes in Computer Science*, pp. 1–18, 2022b.

Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M. Alvarez. FB-OCC: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023b.

Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4105–4114, 2020a.

Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative perception via learnable handshake communication. In *IEEE International Conference on Robotics and Automation*, pp. 6876–6883, 2020b.

Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3d object detection in presence of pose errors. In *IEEE International Conference on Robotics and Automation*, pp. 4812–4818, 2023.

Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Yanfeng Wang, and Siheng Chen. An extensible framework for open heterogeneous collaborative perception. In *International Conference on Learning Representations*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. RenderOcc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *IEEE International Conference on Robotics and Automation*, pp. 12404–12411, 2024.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 815–824, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. LMSCNet: Lightweight multiscale 3d semantic completion. In *International Conference on 3D Vision*, pp. 111–119, 2020.

Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17996–18006, 2024.

Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1746–1754, 2017.

Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3d instance segmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68367–68390, 2023.

Qifan Tan, Wenzhuo Liu, Han Bi, Lening Wang, Lei Yang, Yicheng Qiao, Zhuo Zhao, Yanhuan Jiang, Qiannan Guo, Huaping Liu, Zhiwei Li, and Cheng Qiu. SAMOccNet: Refined SAM-based surrounding semantic occupancy perception for autonomous driving. *Neurocomputing*, 650: 130918, 2025.

Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. SparseOcc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15035–15044, 2024.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A large-scale 3d occupancy prediction benchmark for autonomous driving. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64318–64330, 2023.

Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision*, volume 12347 of *Lecture Notes in Computer Science*, pp. 605–621, 2020.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. OpenOccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17850–17859, 2023.

Yuping Wang, Xiangyu Huang, Xiaokang Sun, Mingxuan Yan, Shuo Xing, Zhengzhong Tu, and Jiachen Li. UniOcc: A unified benchmark for occupancy forecasting and prediction in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25560–25570, 2025.

Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. PanoOcc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17158–17168, 2024.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21729–21740, 2023.

Hanlin Wu, Pengfei Lin, Ehsan Javanmardi, Naren Bao, Bo Qian, Hao Si, and Manabu Tsukada. A synthetic benchmark for collaborative 3d semantic occupancy prediction in v2x-enabled autonomous driving. *arXiv preprint arXiv:2506.17004*, 2025.

Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. SCPNet: Semantic scene completion on point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17642–17651, 2023.

Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, Li Jin, Mingyue Lei, Zhaoyang Ma, Zihang He, Haoxuan Ma, Yunshuang Yuan, Yingqian Zhao, and Jiaqi Ma. V2X-Real: A large-scale dataset for vehicle-to-everything cooperative perception. In *European Conference on Computer Vision*, volume 15110 of *Lecture Notes in Computer Science*, pp. 455–470, 2024.

Huaiyuan Xu, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. A survey on occupancy perception for autonomous driving: The information fusion perspective. *Information Fusion*, 114: 102671, 2025a.

Mutian Xu, Xingyilang Yin, Lingteng Qiu, Yang Liu, Xin Tong, and Xiaoguang Han. SAMPro3D: Locating SAM prompts in 3d for zero-shot instance segmentation. In *International Conference on 3D Vision*, pp. 1222–1232, 2025b.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In *European Conference on Computer Vision*, volume 13699 of *Lecture Notes in Computer Science*, pp. 107–124, 2022a.

Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *IEEE International Conference on Robotics and Automation*, pp. 2583–2589, 2022b.

Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. In *Proceedings of the Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 989–1000, 2023a.

Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, and Jiaqi Ma. V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13712–13722, 2023b.

Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. SAM3D: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21361–21370, 2022.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9433–9443, 2023.

A Additional Comparisons

Table A1 reports final 3D semantic occupancy IoU and mIoU under the up-to-seven-agent setting.

Table A1: Semantic occupancy on Semantic-OPV2V with up to seven agents. IoU is the class-frequency-weighted mean over 12 non-empty classes; mIoU is their arithmetic mean. Scores are percentages; † marks published reference results.

Method	IoU ↑	mIoU ↑
CoHFF† (Song et al., 2024)	50.46	34.16
VOGS-CP† (Chen et al., 2026)	72.44	37.44
VQSOP† (Li et al., 2026)	73.79	41.54
OccWeave (Voxel)	56.67	38.51
OccWeave (Gaussian)	73.13	39.83

Table A2 provides the full BEV semantic segmentation comparison. Predictions are projected along the height axis and evaluated under two-agent and up-to-seven-agent settings.

Table A2: BEV semantic segmentation IoU (%) after height-axis projection. Others groups the nine classes excluding Vehicle, Road, and Sidewalk. Agent counts include the ego vehicle.

Method	# Agents	Vehicle	Road	Others
CoBEVT (Xu et al., 2023a)	2	46.13	52.41	–
CoHFF (Song et al., 2024)	2	47.40	63.36	40.27
VOGS-CP (Chen et al., 2026)	2	70.25	82.69	79.37
VQSOP (Li et al., 2026)	2	73.23	85.05	79.42
OccWeave (Voxel)	2	52.73	74.71	66.95
OccWeave (Gaussian)	2	70.12	87.28	78.55
CoBEVT (Xu et al., 2023a)	Up to 7	60.40	63.00	–
CoHFF (Song et al., 2024)	Up to 7	64.44	57.28	45.89
VOGS-CP (Chen et al., 2026)	Up to 7	75.30	84.96	80.19
VQSOP (Li et al., 2026)	Up to 7	77.48	87.04	81.10
OccWeave (Voxel)	Up to 7	71.84	76.55	65.04
OccWeave (Gaussian)	Up to 7	76.35	90.27	79.25

B Structural Prior Comparison

We compare six prior sources in the voxel implementation to examine the contribution of prior-induced refinement and the value of external pretrained features relative to task-backbone features (Table A3). The tri-plane-to-voxel architecture, task-feature interface, and SCRF design are shared across the settings. SAM supplies the prior features used in the main experiments.

None disables prior-induced refinement by setting its residual strictly to zero, while retaining SCRF. The refined features therefore equal the original task features, and the cross-agent discrepancy is computed from those unchanged features. This control tests cooperative source scoring with prior-induced refinement disabled. In the pre-refinement scoring ablation in Table 3, the fusion values retain SAM-based refinement; here, both scoring and fusion operate without that refinement. Backbone feat. reuses task-network features through the prior adapter, spatial association, and gated residual pathway, testing whether the existing visual representation suffices for this additional refinement.

Edge prior uses OpenCV Canny with fixed preprocessing and thresholds to construct a single-channel edge-density map. Backbone feat. uses the 1,024-channel C4 spatial features of the task network's ResNet101-DCNv2, before the FPN. CLIP feat. uses the 768-channel spatial patch tokens from the

Table A3: Prior-source comparison in the voxel implementation. None disables prior-induced refinement; Backbone feat. reuses task-network features. All settings retain SCRF.

Prior source	Overall		Hard regions mIoU	
	IoU	mIoU	Ego-invis.	Boundary
None	50.33	33.97	19.75	18.79
Edge prior	53.40	34.07	34.22	30.72
Backbone feat.	49.63	33.51	20.15	19.39
CLIP feat.	55.01	35.98	34.64	25.38
DINOv2 feat.	56.38	37.25	35.14	31.87
SAM feat.	56.67	38.51	35.35	32.06

Scores are percentages. Overall columns report semantic occupancy IoU and mIoU; Ego-invis. and Boundary report mIoU on ego-invisible non-empty voxels and class-boundary bands, respectively. None sets the prior-induced residual strictly to zero. Each feature source uses its own trained adapter to the shared task interface.

final visual Transformer of OpenAI CLIP ViT-B/16 (Radford et al., 2021). DINOv2 feat. uses the 768-channel normalized spatial patch tokens of Meta DINOv2 ViT-B/14 without register tokens (Oquab et al., 2024). SAM feat. uses the 256-channel dense neck output of the SAM 1 ViT-B image encoder (Kirillov et al., 2023). All feature sources retain spatial information for association with local 3D features.

Prior extractors remain frozen, and a separate adapter is trained for each source to map its channels to the common task interface before SPFR and SCRF. The sources are associated with the same camera fields of view and 3D query locations. All settings use the same data, training budget, and checkpoint-selection rule. Table A3 shows that reusing backbone features yields lower overall mIoU than None (33.51% versus 33.97%), so the additional refinement pathway alone does not ensure an improvement. Edge cues substantially improve ego-invisible and boundary mIoU, while overall mIoU remains close to None, indicating that their benefit is concentrated in the evaluated spatial subsets. DINOv2 and SAM improve both overall and regional performance, with SAM achieving the highest scores across all four reported metrics among the compared sources. Its overall mIoU exceeds DINOv2 by 1.26 percentage points, while their regional scores are closer. These results support using SAM features for the main implementation and show that the choice of prior representation affects the benefit of refinement.

C Structure-Sensitive Evaluation

Table A4 evaluates the voxel and Gaussian implementations on two spatial subsets and a thin-structure class group. The corresponding baselines are CoHFF and VOGS-CP, respectively. Ego-invisible non-empty voxels are non-empty in the cooperative label and empty in the ego-only label. Class-boundary bands are constructed from cooperative ground truth by marking both voxels of each six-connected, non-empty pair with different semantic labels, dilating this mask by one six-neighbor step, and retaining non-empty voxels. Metrics for these two subsets are computed within their respective masks. The thin-structure group summarizes the global class IoUs for Fence, Pole, Guard rail, and Traffic sign: IoU is their support-weighted mean, and mIoU is their arithmetic mean.

Table A4: Spatial-subset and thin-structure class-group analysis of the voxel and Gaussian implementations.

Variant	Evaluation subset	IoU ↑			mIoU ↑		
		Baseline	OccWeave	Δ	Baseline	OccWeave	Δ
Voxel	All non-empty	50.46	**56.67**	+6.21	34.16	**38.51**	+4.35
	Ego-invisible non-empty	43.35	**59.71**	+16.36	19.42	**35.35**	+15.93
	Class-boundary band	35.50	**46.81**	+11.31	19.00	**32.06**	+13.06
	Thin-structure classes[‡]	30.82	**41.79**	+10.97	22.21	**27.69**	+5.48
Gaussian	All non-empty	72.44	**73.13**	+0.69	37.44	**39.83**	+2.39
	Ego-invisible non-empty	71.93	**73.64**	+1.71	46.10	**48.51**	+2.41
	Class-boundary band	71.96	**73.00**	+1.04	46.09	**47.71**	+1.62
	Thin-structure classes	23.52	**28.28**	+4.75	20.54	**25.37**	+4.83

Scores are percentages; Δ denotes improvement in percentage points. Bold marks the better paired result. Ego-invisible and boundary subsets use spatial masks. [‡]The thin-structure group summarizes global class IoUs for Fence, Pole, Guard rail, and Traffic sign.